

# NIST Research Data Framework (RDaF)

Dr. Robert Hanisch

Director, Office of Data and Informatics (ODI)

Material Measurement Laboratory

US National Institute of Standards and Technology



## About NIST and ODI

- The National Institute of Standards and Technology is a federal agency under the US Department of Commerce
  - Known as the National Bureau of Standards until 1988, originally founded in 1901
- Non-regulatory
- State of the art in measurement science and technology
- US National Metrology Institute, amongst network of 103 NMIs globally organized under the Bureau International des Poids et Mesures (BIPM, or International Bureau of Weights and Measures), Paris
- ~5,000 staff at NIST (Gaithersburg, Maryland headquarters; Boulder, Colorado; Charleston, South Carolina; Brookhaven National Laboratory)
- 6 major research laboratories
  - Material Measurement Laboratory
    - Office of Data and Informatics (15 people)



#### What is a Research Data Framework?

- A map of the research data space: who, what, where, why, when?
- A dynamic guide for the various stakeholders in research data to understand best practices for research data management and dissemination
- A resource for understanding costs, benefits, and risks associated with research data management
- A consensus document based on inputs and conversations amongst the stakeholders in research data



## Why a Research Data Framework?

- Research data ecosystem is very complex!
  - Lots of players, various funding models and sustainability plans
  - How long should data be kept?
  - How should data quality be assessed?
  - How do we measure the value of research data?





#### Big Data Landscape 2016 (Version 3.0) Infrastructure **Analytics Applications** Analyst Analytics **Data Science** Human Hadoop Hadoop in Sales & Marketing Visualization **Customer Service** Spark **Cluster Services** Legal Platforms **Platforms** Capital the Cloud **Platforms** On-Premise RADIUS' Gainsight + ableau ◆ MEDALLIA amazon context relevant databricks **Q** Palantir Sbloomreach Zeta RAVEL Microsoft CONTINUUM & DataRobot Google Cit gild LEVERSTRING livefyre Coogle Cloud Platform JUDICATA AYASDI guavus CLARABRIDG MAPR Pivotal Qlik @ looker blue yonder Lattice GridGain 4 IBM InfoSphere **docker** CLICKFOX Quid enigma Datameer MODE plotly Roamb @kahuna √infer SAILTHRU STELLAService MESOSPHERE textic SISENSE TOOMDATA IBM InfoSphere **TACHYON** Core OS pepperdata Digital Reasoning Bottlenose. NG\*DATA Preac persado AVISO Ósense DOMINO Sense @Brevia entelo bluedata jethro 🙆 altiscale 🔲 bale ORBITAL INSIGHT inter ana ŷhat ALGORITHMIA QUANTIFIND ACTIONIC Stack IQ CHARTIO PREM®NITION hi fuse|machines .ENGAGIO appuri Wise.io Statistical Log Analytics Social **NewSQL Databases BI Platforms NoSQL Databases** Ad Optimization Security Vertical Al Analytics Computing Power BI mazon amazon Google Cloud Platfor Clustrix Pivotal splunk> Hootsuite □ YLANCE Applications AppNexus MediaMath ORACLE **S**sas /// paradigm4 **sumologic** CounterTack cybereason Microsoft Azure MarkLogic NUODB NETBASE criteo. Threat Metrix. splice MACHINE kıbana **DATASIFT ₹** birst SPSS OpenX ∞rocketfuel mongoDB DATASTAX GoodData AREA 1 SentinelOne Recorded Future Guardian tracx bitly CLOUD PHYSICS Integral () the Trade Desk O Clara **∢EROSPIKE** citusdata platfora MATLAB **synthesio** dstillery deepdb Trafodion Cockroach LABS SequoiaDB redislabs @ influxdata loggly **KASIST** (D) atscale **TFORT**SCALE \*siftscience SKeybase feedzai #SICNIFYD lumiata DataXU Appier MOA Graph MPP Data Cloud EDW Real-Time Machine Learning Speech & NLP Horizontal AI **Databases** Databases Transformation Integration BM Watson amazon Finance Publisher Govt / Regulation amazon informatica sentient alteryx Google Coud ✓ III LendingClub meo4j H<sub>2</sub>O Tools METAMARKETS NUANCE W Socrata VERTICA vicarious Microsoft Azure talend outbrain OnDeck> ...Kreditech MuleSoft Sstriim Dato Pivotal TO OPENGOV N NETEZZA TRIFACTA noro 🕞 🕰 Numenta Tab<sub>2</sub> antinance Lendup 🧖 Kabbage snaplogic api.ai tamr 👡 🛚 Oction quantcast FN FiscalNote tidemark. INSIKT **BedrockData** Descartes clarifai Orient DB' DATATORREN MindMeld ... StreamSets Chartbeat Z UOra Dataminr A Lenddo dataArtisans EXASOL Odremio PREDPOL Infoworks M Alation xplenty IDIBON (2) a yieldbot mark43 **ISENTIUM** Web / Mobile Data Services For Business Storage Crowd-Search Management Security App Dev """ OpenDataSoft sentient Quantopian Yieldmo TANIUM" Analysts / Commerce / Monitoring sourcing UO OPERA apigee Google Analytics New Relic. OrigamiLogic 35 EXALEAD Education/ Life Sciences Industries Counsyl PATHWAY GENOMICS mixpanel C CODE 42 Microsoft Azu Lucidworks amazon actifio EXL OP@WER eHarmony Learning ClearStory Data Gravity panasas RJMetrics BLUECORI elastic ThoughtSpot Numerify splunk> RetailNext XX Recombine © CipherCloud AMPLITUDE ( granify **Typesafe** nimblest DATA SCIENCE CIRRO **™**VECTR∧ M∧∧N∧ Ø swiftype KYRUUS FLATIRON STITCH FIX Clever соно sumal Airtable ©@⊘@zymergen HealthTap® import (io retention custora TACHYUS Seeq FarmLogs Frocana DRIVEN Qumulo @eclara METABIOTA ZEPHYR Cross-Infrastructure/Analytics PANORAMA HowGood celect (6) SIGHT Ginger.io \* transcriptic Glow knowre Cenlific AiCure A At statmuse B@XEVER Tamazon Google Microsoft The SAP SSAS III M AMMONIN VERTICA VMWARE TIBO TERADATA ORACLE I NetApp **Open Source** Framework Query / Data Flow Data Access Coordination Stat Tools Machine Learning % Real-Time Search Security HBASE mongoDF accumulo Apache SINGA MADlib. Spork Spork Apache Ranger talend 🚾 🕮 🖪 cassandra YARN MESOS ScalaLab Apache Zookeeper Caffe VELES FeatureFu CNTK TensorFlo Visualization CouchDB %riak # OPENTSD Solr Spark SLAMDATA APACHE \*\* TACHYON - druid DIMSUM Jupyter DL4J Apache Ambari Data Sources & APIs Incubators & Schools Location / People / Entities Health Financial & Economic Data Air / Space / Sea Other JAWBONE GARMIN. PLURALSIGHT Bloomberg D | DOW JONES acxiem Experian Epsilon InsideView ▲ spire DataCamp Y DLEE PREMISE CAPITAL IQ GARMIN COURSQUOR STREETLINE @esri practice fusion ... fitbit helium samsara A DataElite panjiva Withings VALIDIC netatm Crimson Hexagon 🌎 CARTODB 📑 factual. quandl xignite CBINSIGHTS The Data Incubator METIS **©**kinsa Human API StockTwits @estimize PLAID Airware Tone Deploy CIRCULATE @ placemeter ĐATA.GOV Last Updated 3/23/2016 © Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap) FIRSTMARK

ATORY

NST

## Why a Research Data Framework?

And how do we leverage research data to address

global challenges?



#### **RDaF Benefits**

- Increase research integrity with quality data and improved transparency of the research process
- Reduce costs and maximize efficiency by establishing best practices for data management
- Guide risk management and reduction through assessment of risk positions and roadmaps for improvement
- Increase scientific discovery and innovation with the FAIR principles (Findable, Accessible, Interoperable, Reusable) for better utilization of data



#### National and International Need

- Data is proliferating at an exponential rate
- Data management is complex and confusing
- Mismanaged data has dire social and economic consequences, including loss of global leadership in critical technical fields
- The U.S. needs a coordinated effort to establish a research data infrastructure, but research data are global in nature so international collaboration / coordination is necessary
- NIST is well-positioned to lead the project; our business is consensus building through being a neutral convener of diverse communities



#### Stakeholders

- Government agencies
- Universities and research libraries
- Data repositories
- Scholarly publishers
- Professional societies
- National and international collaboration organizations (e.g., CENDI, BRDI, CODATA, RDA, WDS, GO-FAIR)
- Standards bodies
- Funders (public and private)
- Industry and the private sector
- Researchers
- General public



#### **Process**

- Pilot program to provide an overall guide to the actors and stakeholders in the research data space
- NIST Cybersecurity Framework is the model
- Community consensus, not NIST imposition
- If I am a \_\_\_\_\_, then I need to know \_\_\_\_\_.
- Initial scoping workshop held in December 2019 at NIST NCCoE
  - 50 invited participants representing stakeholders, both US and international





## NIST RDaF Workshop



## Research Data Framework

Robert Hanisch

Director, Office of Data and Informatics

Material Measurement Laboratory

National Institute of Standards and Technology





Bonnie Carroll
Founder & CDO,
Information
International
Associates
Secretary General,
CODATA

RDaF Steering Group



Laura Biven, DOE



Mercé Crosas, Harvard





Josh Greenberg, Sloan Hilary Hanahoe, RDA





Heather Joseph, SPARC Barend Mons, CODATA and GO-FAIR



Beth Plale, NSF



Anita de Waard, Elsevier

# Opening Keynotes



Walter Copan, NIST Director



Jim St. Pierre, Deputy Director Information Technology Laboratory

## Workshop Summary

- Why an RDaF? An increasing complex data ecosystem with voracious artificial intelligence drivers for trustworthy data requires systematic, intentional data asset management.
- **Purpose:** To optimize use and value of strategic data assets with a coherent data management strategy.
- **Scope:** Covers management of research data created and/or used by any organization. Provides a customizable framework to select and evaluate maturity on key data management factors for each organization.

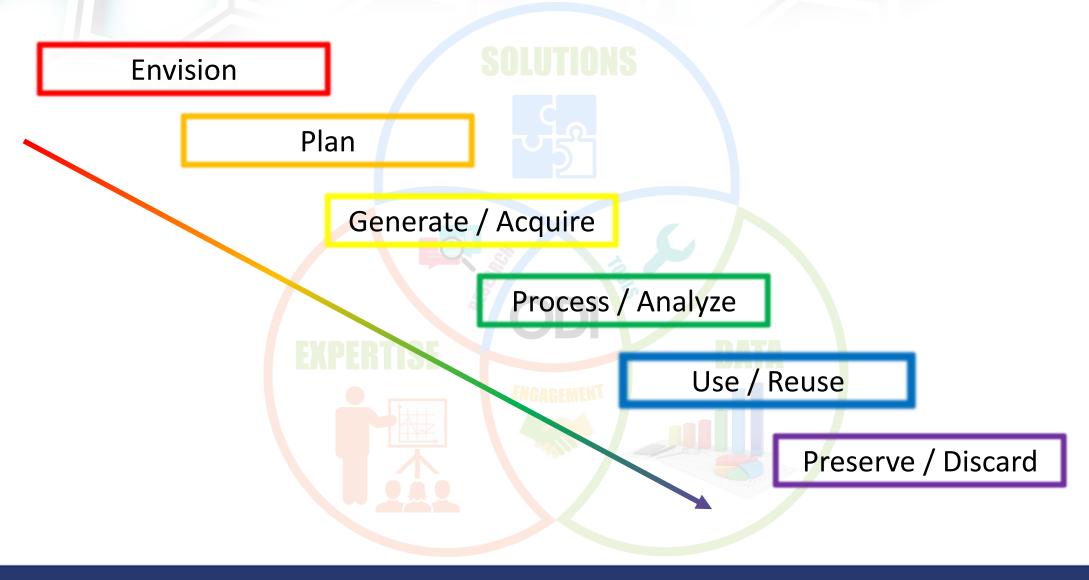


## Workshop Summary (cont'd)

- **Status:** Confirmed support by government agencies, academic organizations, private sector companies, not-for-profit organizations, and international stakeholders.
- **Next Steps:** Management commitment to complete the scoping, pilot testing, and community building for the Framework.
  - Proposed pilots
    - Materials science
    - Universities and research libraries (AAU, APLU, ARL)

Will need cooperation across government to move fully forward with the Framework.

## RDaF Structure Based on "Functions"





**Function Category** Subcategory 1) Envision Data governance Data vision, data policy Data management organization Data quality, privacy, ethics Communication, interactions Community Cross-domain engagement Data culture FAIR principles Value of data Roles and responsibilities Value of data professionals Reward structure Incentives for sharing and re-use

**Function** Category Subcategory Cost-benefit analysis 2) Plan Costs Costs by data lifecycle stage **Funding** Direct, overhead, mixed, other Data objects Data (experimental, simulation) Software, instruments Publications, presentations DMPs (intent, update) Data management Formats, standards planning



**Function** Subcategory Category 3) Generate / In-house, experiment or simulation Sources Collected from external sources Acquire Instruments and their metadata Experiment Measurement protocols Data capture and recording Simulation Commercial or custom software Metadata capture and recording External sources Identification, provenance Metadata harvesting

Data formats

Standards development and/or adoption

**Function** Subcategory Category 4) Process / Origin, version, time-stamp Provenance Data copied or derived from other data **Analyze** Data architecture Design, security, configuration management Hosting and storage On-premise or Cloud Software Commercial or custom software Versions Stability, resilience, adaptability, maintenance Workflows, ELNs, LIMs Publishing, curation Processes, tools, stewardship Metadata



**Function** Subcategory Category 5) Use / Reuse Legal and licenses Ownership, IP, rights and restrictions Agreements, permissions Citation expectations Internal, external Data access **APIs** Downloads vs. visiting Analysis tools AI/ML Performance Usage tracking, citation *Impact* 

**Function** Subcategory Category 6) Preserve / Sustainability Longevity requirements **Discard** Who pays? Orphan data sets Media and media migration Preservation Back-up Repositories (domain, institutional, general) Migration between organizations Retention and disposition Decision processes End-of-life (dark archives, deaccession, gravestones)

#### Status

 Briefed OSTP Subcommittee on Open Science and OSTP Director Kelvin Droegemeier (03/26/2020)

- Discussed pilot concept with AAU, APLU, ARL
- Developed roadmap and structure, vetted with Steering Group
- Preparing to brief NIST upper management
- Will seek ~\$450k to fund two pilots
  - NIST plus other agencies, either \$\$ or in-kind support
  - Professional societies
  - Scholarly publishing community



#### **NIST Frameworks**

Framework for Improving Critical Infrastructure Cybersecurity:

https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf

NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, September 6, 2019 (Preliminary Draft)

https://www.nist.gov/system/files/documents/2019/09/09/nist privacy framework preliminary draft.pdf

NIST Big Data Interoperability Framework: Volume 1, Definitions October 2019 Version 3

https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1r2.pdf



## **RDaF Summary**

- Successful in building community interest and engagement
  - Diverse stakeholders
  - National and international
- Challenges
  - Resources, including our own time within ODI/MML
  - Timeliness: the research data ecosystem is changing rapidly. How to keep pace and assure ongoing updates?
  - Controlling scope and scale
- Strategy for moving forward
  - Confirm NIST support with upper management
  - Start with pilot projects in order to validate approach and re-tune as necessary
  - Collaborate with other federal agencies, professional societies, scholarly publishing community, etc., to garner the necessary resources and take advantage of work in progress



## Contact

#### **SOLUTIONS**

Robert Hanisch

robert.hanisch@nist.gov

https://nist.gov/people/robert-hanisch



